Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 21 (2007) 609–619

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Learning model order from labeled and unlabeled data for partially supervised classification, with application to word sense disambiguation

Zheng-Yu Niu [a,*], Dong-Hong Ji [a], Chew Lim Tan [b]

[a] Institute for Infocomm Research, Mail Box B023, 21 Heng Mui Keng Terrace, Singapore 119613, Singapore
[b] Department of Computer Science, National University of Singapore, 3 Science Drive 2, Singapore 117543, Singapore

Received 6 April 2006; received in revised form 2 February 2007; accepted 28 February 2007
Available online 12 March 2007

Abstract

Previous partially supervised classification methods can partition unlabeled data into positive examples and negative examples for a given class by learning from positive labeled examples and unlabeled examples, but they cannot further group the negative examples into meaningful clusters even if there are many different classes in the negative examples. Here we proposed an automatic method to obtain a natural partitioning of mixed data (labeled data + unlabeled data) by maximizing a stability criterion defined on classification results from an extended label propagation algorithm over all the possible values of model order (or the number of classes) in mixed data. Our experimental results on benchmark corpora for word sense disambiguation task indicate that this model order identification algorithm with the extended label propagation algorithm as the base classifier outperforms SVM, a one-class partially supervised classification algorithm, and the model order identification algorithm with semi-supervised k-means clustering as the base classifier when labeled data is incomplete.

Keywords: Word sense disambiguation; Partially supervised classification; Semi-supervised clustering

1. Introduction

Partially supervised classification algorithms (Denis et al., 2002; Liu et al., 2003; Manevitz and Yousef, 2001; Yu et al., 2002) try to learn from positive examples and unlabeled examples for a given class. After the learning procedure, unlabeled examples will be classified into positive examples and negative examples. We can see that these methods always assume the occurrence of negative examples in unlabeled data. Moreover, they cannot group negative examples into meaningful clusters even if there are many different classes in the negative examples. But, for some classification tasks, it is better to further group negative examples into meaningful clusters if negative data includes the instances from many different classes.

[*] Corresponding author. Tel.: +65 93612080.
E-mail addresses: zniu@i2r.a-star.edu.sg (Z.-Y. Niu), dhji@i2r.a-star.edu.sg (D.-H. Ji), tancl@comp.nus.edu.sg (C.L. Tan).

0885-2308/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2007.02.001

Here we use the word sense disambiguation problem as an example to show the motivation of our work. We define the problem of partially supervised word sense disambiguation as the disambiguation of senses of occurrences of a target word in untagged texts when given incomplete tagged corpus[1] for this target word. Partially supervised sense disambiguation may help enrich manually compiled lexicons by discovering new senses from untagged corpora.

Word sense disambiguation (WSD) can be defined as associating a target word in a text or discourse with a definition or meaning. Many corpus based methods have been proposed to deal with the sense disambiguation problem when given a definition for each possible sense of a target word or a tagged corpus with the instances of each possible sense, e.g. supervised sense disambiguation methods, and semi-supervised sense disambiguation methods.

Supervised sense disambiguation methods rely on the information from previously sense tagged corpora to determine the senses of words in unseen texts (Leacock et al., 1998; Lee and Ng, 2002; Pedersen, 2000).

Semi-supervised methods for sense disambiguation are characterized in terms of exploiting unlabeled data in the learning procedure with the requirement of predefined sense inventories for target words. Some methods were proposed to exploit bilingual resources, e.g. aligned parallel corpora, untagged monolingual corpora in two languages Brown et al., 1991; Dagan and Itai, 1994; Diab and Resnik, 2002; Li and Li, 2004; Ng et al., 2003. Another research line is to automatically generate monolingual sense tagged corpora without reference to the second language corpora (Hearst, 1991; Karov and Edelman, 1998; Mihalcea and Moldovan, 1999; Mihalcea, 2004; Niu et al., 2005; Park et al., 2000; Pham et al., 2005; Yarowsky, 1995), e.g. bootstrapping.

Bootstrapping (or self-training) is a general scheme for minimizing the requirement of manually tagged corpus, which was proposed for sense disambiguation by Hearst (1991).

Hearst's bootstrapping method was improved by Yarowsky (1995) in two aspects: (a) manually identify collocations for word senses to generate initial labeled data; (b) exploit a redundant view (one sense per discourse property) to filter or augment sense tagged examples in the bootstrapping process.

Some efforts were devoted to improve the base classifier in the bootstrapping process: Park et al. (2000) used committee learning algorithm as the base classifier, while Mihalcea (2004) introduced a combination of majority voting with bootstrapping or co-training.

Mihalcea and Moldovan (1999) did some work on using web data to obtain sense tagged corpora. They used the information from WordNet to formulate queries consisting of synonyms or definitions of word senses, and obtained additional training data for word senses from Internet using existing search engines.

Recently, Pham et al. (2005) described an application of four semi-supervised learning algorithms for WSD, including basic co-training, smoothed co-training, spectral graph transduction (SGT), and a variant of SGT (SGT + co-training). Niu et al. (2005) evaluated a graph based semi-supervised learning algorithm on SENSEVAL-3 data for WSD.

Some observations can be made on previous supervised and semi-supervised methods. They always rely on hand-crafted lexicons (e.g. WordNet) as sense inventories. But these resources may miss domain-specific senses, which leads to incomplete sense tagged corpora. Therefore, sense taggers trained on the incomplete tagged corpora may misclassify some instances if the senses of these instances are not defined in sense inventories. For example, one performs word sense disambiguation in information technology related texts using WordNet[2] as sense inventory. When disambiguating word "boot" in phrase "boot sector", a sense tagger will assign this instance with one of the senses of "boot" listed in WordNet. But the correct sense "loading operating system into memory" is not included in WordNet. Therefore, this instance may be associated with an incorrect sense. We can use partially supervised classification algorithms to retrieve instances from untagged corpus for each sense appearing in the incomplete tagged corpus. However, these methods cannot further group the negative examples (instances not belonging to any sense in the incomplete tagged corpus) into meaningful clusters even if there are many different senses in the negative examples.

This partially supervised sense disambiguation problem may be generalized as a partially supervised classification problem by treating the incomplete sense tagged corpus of a target word (consisting of the contexts

[1] "Incomplete tagged corpus" means that tagged corpus does not include the instances of some senses for the target word, while these senses may occur in untagged texts.

[2] Online version of WordNet is available at http://wordnet.princeton.edu/cgi-bin/webwn2.0

of occurrences of the target word with correct sense tags) as labeled data, the untagged corpus of the target word as unlabeled data, each possible sense of the target word as a class, and the number of senses (or sense number) of the target word in mixed data (labeled data + unlabeled data) as model order (or cluster number, class number). Given labeled data and unlabeled data, we are interested in (1) labeling the instances in unlabeled data with class labels appearing in labeled data; (2) trying to find each possible new class[3] from unlabeled data.

For discovering each possible new class in unlabeled data, we propose an automatic method to estimate the model order of mixed data by maximizing a stability criterion defined on the classification results from an extended label propagation algorithm over all the possible values of model order. After model order identification, we can obtain a partitioning of the mixed data with the optimal number of clusters, where each cluster consists of similar examples from mixed data, and it satisfies two constraints: labeled examples with the same class label will stay in the same cluster, labeled examples with different class labels will stay in different clusters. If the estimated model order (or cluster number) of mixed data is equal to the number of classes in labeled data, then there is no new class in unlabeled data. Otherwise new classes will be represented by the clusters in which there is no instance from labeled data.

This paper is organized as follows. First, we will present a model order identification algorithm with an extended label propagation algorithm as the base classifier for partially supervised classification in Section 2. Then we will provide experimental results of this algorithm on SENSEVAL-3 data for WSD in Section 3. We will summarize related work on partially supervised classification in Section 4. Finally we will conclude our work and suggest possible improvements in Section 5.

2. Model order identification for partially supervised classification

We perform partially supervised classification by following steps:

(1) Estimate the optimal value of model order of mixed data by maximizing a stability criterion defined on classification results from an extended label propagation algorithm (to be presented in Section 2.1) over all possible values of model order in mixed data. The stability criterion assesses the agreement between classification result on full mixed data and that on sampled mixed data. An extended label propagation algorithm is used to classify the full or sampled mixed data into a given number of clusters before the stability assessment. We will provide the details of the model order identification procedure and the stability criterion in Section 2.2.

(2) After model order identification (or cluster number estimation), we can obtain a partitioning of mixed data with the estimated number of clusters using the extended label propagation algorithm, where each cluster consists of similar examples from mixed data, and it satisfies two constraints: labeled examples with the same class label will stay in the same cluster, labeled examples with different class labels will stay in different clusters. In fact, this classification process has been performed on mixed data in the order identification procedure.

2.1. An extended label propagation algorithm

Let $X_{\mathrm{L+U}} = \{x_i\}_{i=1}^{n}$ be a set of labeled and unlabeled examples (mixed data) for a target word, where x_i represents the ith example, and n is the total number of examples. Let $S_{\mathrm{L}} = \{s_j\}_{j=1}^{c}$ denote the class label set in X_{L}, where X_{L} consists of the first l examples x_g $(1 \leqslant g \leqslant l)$ that are labeled as y_g $(y_g \in S_{\mathrm{L}})$. Let X_{U} denote other u $(l + u = n)$ examples $x_h(l+1 \leqslant h \leqslant n)$ that are unlabeled.

Let $Y^0_{X_{\mathrm{L+U}}} \in N^{|X_{\mathrm{L+U}}|\times|S_{\mathrm{L}}|}$ represent initial soft labels attached to labeled examples, where $Y^0_{X_{\mathrm{L+U}},ij} = 1$ if y_i is s_j and 0 otherwise. Let $Y^0_{X_{\mathrm{L}}}$ be the top l rows of $Y^0_{X_{\mathrm{L+U}}}$ and $Y^0_{X_{\mathrm{U}}}$ be the remaining u rows. $Y^0_{X_{\mathrm{L}}}$ is consistent with the labeling in labeled data, and the initialization of $Y^0_{X_{\mathrm{U}}}$ can be arbitrary.

[3] "New class" is the class that does not appear in labeled data.

Let k denote the possible value of model order in mixed data X_{L+U}, and k_{X_L} be the number of classes in initial tagged data X_L. Note that $k_{X_L} = |S_L|$, and $k \geqslant k_{X_L}$.

The classification algorithm in the order identification process should be able to accept labeled data D_L[4], unlabeled data D_U[5] and model order k as input, and assign a class label or a cluster index to each instance in D_U as output. Traditional supervised or semi-supervised algorithms (e.g. SVM, label propagation algorithm (Zhu and Ghahramani, 2002)) cannot classify the examples in D_U into k clusters if $k > k_{X_L}$. The semi-supervised k-means clustering algorithm (Wagstaff et al., 2001) may be used to perform clustering analysis on mixed data, but its efficiency is a problem for clustering analysis on a very large dataset since multiple restarts are usually required to avoid local optima and multiple iterations will be run in each clustering process for optimizing a clustering solution.

In this work, we propose an alternative method, an extended label propagation algorithm (ELP), which can classify the examples in D_U into k clusters. If the value of k is equal to k_{X_L}, then ELP is identical with the plain label propagation algorithm (LP) (Zhu and Ghahramani, 2002). Otherwise, if the value of k is greater than k_{X_L}, we will perform classification by following steps:

(1) estimate the size of the dataset of each new class as $\text{size}_{\text{new\_class}}$ by identifying the examples of new classes using the "Spy" technique[6] and assuming that new classes are equally distributed;
(2) $D'_L = D_L$, $D'_U = D_U$;
(3) remove tagged examples of the mth new class ($k_{X_L} + 1 \leqslant m \leqslant k$) from D'_L[7] and train a classifier on this labeled dataset without the mth class;
(4) the classifier is then used to classify the examples in D'_U;
(5) the least confidently unlabeled point $x_{\text{class\_}m} \in D'_U$, together with its label m, is added to the labeled data $D'_L = D'_L + x_{\text{class\_}m}$, and $D'_U = D'_U - x_{\text{class\_}m}$;
(6) steps (3) to (5) are repeated for each new class till the augmented tagged data set is large enough (here we try to select $\text{size}_{\text{new\_class}}/4$ examples with their sense tags as tagged data for each new class);
(7) use the plain LP algorithm to classify remaining unlabeled data D'_U with D'_L as labeled data.

Table 1 shows the details of this extended label propagation algorithm.

Next we will describe the plain label propagation algorithm.

Define $W_{ij} = \exp(-\frac{d_{ij}^2}{\sigma^2})$ if $i \neq j$ and $W_{ii} = 0$ $(1 \leqslant i,j \leqslant |D_L + D_U|)$, where d_{ij} is the distance (e.g. Euclidean distance) between the example x_i and x_j, and σ is used to control the weight W_{ij}.

Define $|D_L + D_U| \times |D_L + D_U|$ probability transition matrix $T_{ij} = P(j \to i) = \frac{W_{ij}}{\sum_{k=1}^{n} W_{kj}}$, where T_{ij} is the probability to jump from example x_j to example x_i.

Compute the row-normalized matrix $\overline{T}$ by $\overline{T}_{ij} = T_{ij} / \sum_{k=1}^{n} T_{ik}$.

The classification solution is obtained by $Y_{D_U} = (I - \overline{T}_{uu})^{-1} \overline{T}_{ul} Y^0_{D_L}$. I is $|D_U| \times |D_U|$ identity matrix. $\overline{T}_{uu}$ and $\overline{T}_{ul}$ are acquired by splitting matrix $\overline{T}$ after the $|D_L|$th row and the $|D_L|$th column into 4 sub-matrices.

2.2. Model order identification procedure

For achieving the model order identification ability, we use a cluster validation based criterion (Levine and Domany, 2001) to infer the optimal value of model order of X_{L+U}.

[4] D_L may be the dataset X_L or a subset sampled from X_L.

[5] D_U may be the dataset X_U or a subset sampled from X_U.

[6] The "Spy" technique was proposed in Liu et al. (2003). Our re-implementation of this technique consists of three steps: (1) sample a small subset D^s_L with the size $15\% \times |D_L|$ from D_L; (2) train a classifier with tagged data $D_L - D^s_L$; (3) classify D_U and D^s_L, and then select some examples from D_U as the dataset of new classes, which have the classification confidence less than the average of that in D^s_L. Classification confidence of the example x_i is defined as the absolute value of the difference between two maximum values from the ith row in labeling matrix.

[7] Initially there are no tagged examples for the mth class in D'_L. Therefore we do not need to remove tagged examples for this new class, and then directly train a classifier with D'_L.

Table 1
An extended label propagation algorithm

```
    Function: ELP($D_L$, $D_U$, $k$, $Y^0_{D_L+D_U}$)
    Input: labeled examples $D_L$, unlabeled examples $D_U$,
    Model order $k$, initial labeling matrix $Y^0_{D_L+D_U}$;
    Output: the labeling matrix $Y_{D_U}$ for $D_U$;
1   If $k < k_{X_L}$ then
        $Y_{D_U} = \mathrm{NULL}$;
2   Else if $k = k_{X_L}$ then
        Run plain label propagation algorithm on $D_U$ with $Y_{D_U}$ as output;
3   Else then
3.1     Estimate the size of tagged data set of each new class;
3.2     Generate tagged examples from $D_U$ for $(k_{X_L} + 1)$th to $k$th new classes;
3.3     Run the plain label propagation algorithm on $D_U$ with augmented tagged dataset as labeled data;
3.4     $Y_{D_U}$ is the output from plain label propagation algorithm;
    End if
4   Return $Y_{D_U}$
```

Table 2
Stability criterion

```
    Function: CV($X_{L+U}$, $k$, $q$, $Y^0_{X_{L+U}}$)
    Input: data set $X_{L+U}$, model order $k$, and sampling frequency $q$;
    Output: the score of the merit of $k$;
1   Run the extended label propagation algorithm with $X_L$,
    $X_U$, $k$ and $Y^0_{X_{L+U}}$;
2   Construct connectivity matrix $C_k$ based on above classification solution on $X_U$;
3   Use a random predictor $\rho_k$ to assign uniformly drawn labels to each vector in $X_U$;
4   Construct connectivity matrix $C_{\rho_k}$ based on above classification solution on $X_U$;
5   For $\mu = 1$ to $q$ do
5.1     Randomly sample a subset $X^\mu_{L+U}$ with the size $\alpha|X_{L+U}|$ from $X_{L+U}$, $0 < \alpha < 1$;
5.2     Run the extended label propagation algorithm with $X^\mu_L$, $X^\mu_U$, $k$ and $Y^{0\mu}$;
5.3     Construct connectivity matrix $C^\mu_k$ using above classification solution on $X^\mu_U$;
5.4     Use $\rho_k$ to assign uniformly drawn labels to each vector in $X^\mu_U$;
5.5     Construct connectivity matrix $C^\mu_{\rho_k}$ using above classification solution on $X^\mu_U$;
    Endfor
6   Evaluate the merit of $k$ using following formula:
    $M_k = \frac{1}{q}\sum_\mu (M(C^\mu_k, C_k) - M(C^\mu_{\rho_k}, C_{\rho_k}))$, where $M(C^\mu, C)$ is given by Eq. (2);
7   Return $M_k$;
```

The model order identification procedure can be formulated as:

$$\hat{k}_{X_{L+U}} = \mathrm{argmax}_{K_{\min} \leqslant k \leqslant K_{\max}} \{CV(X_{L+U}, k, q, Y^0_{X_{L+U}})\}. \tag{1}$$

$\hat{k}_{X_{L+U}}$ is the estimated model order in X_{L+U}, $K_{\min}$ (or $K_{\max}$) is the minimum (or maximum) value of model order, and k is the possible value of model order in X_{L+U}. Note that $k \geqslant k_{X_L}$. We set $K_{\min} = k_{X_L}$. $K_{\max}$ will be set as a value greater than any possible ground-truth value. CV is a cluster validation based evaluation function, or stability criterion. Table 2 shows the details of this function. We set q, the resampling frequency for estimation of stability score, as 20. α is set as 0.90. The random predictor assigns uniformly distributed class labels to each instance in a given dataset. We run this CV procedure for each value of k. The value of k that maximizes this CV function will be selected as the estimation of model order. At the same time, we can obtain a partitioning of X_{L+U} with $\hat{k}_{X_{L+U}}$ clusters.

The function $M(C^\mu, C)$ in Table 2 is given by Levine and Domany (2001):

$$M(C^\mu, C) = \frac{\sum_{i,j} 1\{C^\mu_{i,j} = C_{i,j} = 1, x_i, x_j \in X^\mu_U\}}{\sum_{i,j} 1\{C_{i,j} = 1, x_i, x_j \in X^\mu_U\}}, \tag{2}$$

where X^μ_U is the untagged data in X^μ_{L+U}, X^μ_{L+U} is a subset with the size $\alpha|X_{L+U}|$ $(0 < \alpha < 1)$ sampled from X_{L+U}, C or C^μ is $|X_U| \times |X_U|$ or $|X^\mu_U| \times |X^\mu_U|$ connectivity matrix based on classification solutions computed on X_U or X_U^μ, respectively. The connectivity matrix C is defined as: $C_{i,j} = 1$ if x_i and x_j belong to the same cluster, otherwise $C_{i,j} = 0$. C^μ is calculated in the same way.

$M(C^\mu, C)$ measures the proportion of example pairs in each cluster computed on X_U that are also assigned into the same cluster by the classification solution on X_U^μ. Clearly, $0 \leqslant M \leqslant 1$. Intuitively, if the value of k is identical with the true value of model order, then classification results on different subsets generated by sampling should be similar with that on full dataset. In the other words, the classification solution with the true model order as input is robust against resampling, which gives rise to a local optimum of $M(C^\mu, C)$.

In this algorithm, we normalize $M(C_k^\mu, C_k)$ by the equation in step 6 of Table 2, which makes our objective function different from the figure of merit (Eq. (2)) proposed in Levine and Domany (2001). The reason to normalize $M(C_k^\mu, C_k)$ is that $M(C_k^\mu, C_k)$ tends to decrease when increasing the value of k (Lange et al., 2002). Therefore for avoiding the bias that the smaller value of k is to be selected as the model order, we use the cluster validity of a random predictor to normalize $M(C_k^\mu, C_k)$.

If $\hat{k}_{X_{L+U}}$ is equal to k_{X_L}, then there is no new sense in X_U. Otherwise ($\hat{k}_{X_{L+U}} > k_{X_L}$) new senses may be represented by clusters in which there is no instance from X_L.

3. Experiments and results

3.1. Experiment design

We evaluated the ELP based model order identification algorithm (implemented by ourselves) on the data in English lexical sample task of SENSEVAL-3 (including all the 57 English words)[8] for WSD, and further empirically compared it with other state of the art classification methods, including SVM[9] (the state of the art method for supervised WSD (Mihalcea et al., 2004)), a one-class partially supervised classification algorithm (Liu et al., 2003)[10], and a semi-supervised k-means clustering based model order identification algorithm (implemented by ourselves).

Given an incomplete tagged corpus for a target word, SVM does not have the ability to find the new senses from untagged corpus. Therefore it labels all the instances in the untagged corpus with sense tags from S_L.

Given a set of positive examples for a class and a set of unlabeled examples, the one-class partially supervised classification algorithm, learning from positive and unlabeled examples (LPU) (Liu et al., 2003), learns a classifier in four steps:

Step 1: Identify a small set of reliable negative examples from unlabeled examples by the use of a classifier.
Step 2: Build a classifier using positive examples and automatically selected negative examples.
Step 3: Iteratively run previous two steps until no unlabeled examples are classified as negative ones or the unlabeled set is null.
Step 4: Select a good classifier from the set of classifiers constructed above.

For comparison, LPU[11] was run to perform classification on X_U for each class in X_L. The label of each instance in X_U was determined by maximizing the classification score from LPU output for each class. If the maximum score of an instance is negative, then this instance will be labeled as a new class. Note that LPU classifies X_{L+U} into $k_{X_L} + 1$ groups in most of cases.

The clustering based partially supervised sense disambiguation algorithm was implemented by replacing ELP with the semi-supervised k-means clustering algorithm (Wagstaff et al., 2001) in the model order identification procedure. The label information in labeled data was used to guide the semi-supervised clustering on X_{L+U}. Firstly, the labeled data may be used to determine initial cluster centroids. If the cluster number is

[8] Available at http://www.senseval.org/senseval3.

[9] We used a linear SVM$^{\text{light}}$, available at http://svmlight.joachims.org/.

[10] Available at http://www.cs.uic.edu/~liub/LPU/LPU-download.html.

[11] The three parameters in LPU were set as follows: "−s1 spy −s2 svm −c1". It means that we used the spy technique for step 1 in LPU, the SVM algorithm for step 2, and selected the first or the last classifier as the final classifier. It is identical with the algorithm "Spy + SVM IS" in Liu et al. (2003).

Table 3
Description of the percentage of official training data used as tagged data when the instances with different sense sets are removed from official training data

	The percentage of official training data used as tagged data (%)
$S_{\text{subset}} = \{s_1\}$	42.8
$S_{\text{subset}} = \{s_2\}$	76.7
$S_{\text{subset}} = \{s_3\}$	89.1
$S_{\text{subset}} = \{s_1, s_2\}$	19.6
$S_{\text{subset}} = \{s_1, s_3\}$	32.0
$S_{\text{subset}} = \{s_2, s_3\}$	65.9

greater than k_{X_L}, the initial centroids of clusters for new classes will be assigned as randomly selected instances. Secondly, in the clustering process, the instances with the same class label will stay in the same cluster, while the instances with different class labels will belong to different clusters. For better clustering solution, this clustering process will be restarted three times. Clustering process will be terminated when clustering solution converges or the number of iteration steps is more than 30. $K_{\min} = k_{X_L} = |S_L|$, $K_{\max} = K_{\min} + m$. m is set as 4.

The data for English lexical samples task in SENSEVAL-3 consists of 7860 examples as official training data, and 3944 examples as official test data for 57 English words. The number of senses of each English word varies from 3 to 11.

We evaluated these four algorithms with different incomplete tagged datasets. Given official training data of word w, we constructed incomplete tagged data X_L by removing the all the tagged instances from official training data that have sense tags from S_{subset}, where S_{subset} is a subset of the ground-truth sense set S for w, and S consists of the sense tags in official training set for w. The removed training data and official test data of w were used as X_U. Note that $S_L = S - S_{\text{subset}}$. Then we ran these four algorithm for each target word w with X_L as tagged data and X_U as untagged data, and evaluated their performance using the accuracy on official test data of all the 57 words.[12] We conducted six experiments for each target word w by setting S_{subset} as $\{s_1\}$, $\{s_2\}$, $\{s_3\}$, $\{s_1, s_2\}$, $\{s_1, s_3\}$, or $\{s_2, s_3\}$, where s_i is the ith most frequent sense of w. S_{subset} cannot be set as $\{s_4\}$ since some words have only three senses. Table 3 lists the percentage of official training data used as tagged data (the number of examples in incomplete tagged data divided by the number of examples in official training data) when we removed the instances with sense tags from S_{subset} for all the 57 words. If $S_{\text{subset}} = \{s_3\}$, then most of sense tagged examples still stay in tagged data. If $S_{\text{subset}} = \{s_1, s_2\}$, then there are very few tagged examples in tagged data. If no instances are removed from official training data, then the value of percentage is 100%.

We used Jensen–Shannon (JS) divergence (Lin, 1991) as distance measure for semi-supervised clustering and ELP, since plain LP with JS divergence achieves better performance than that with cosine similarity on SENSEVAL-3 data (Niu et al., 2005).

For the plain LP algorithm in ELP, we constructed connected graphs as follows: two instances u,v will be connected by an edge if u is among v's 10 nearest neighbors, or if v is among u's 10 nearest neighbors as measured by cosine or JS distance measure (following (Zhu and Ghahramani, 2002)).

We used three types of features to capture the information in all the contextual sentences of target words in SENSEVAL-3 data for all the four algorithms: part-of-speech of neighboring words with position information, words in topical context without position information (after removing stop words), and local collocations (as same as the feature set used in Lee and Ng (2002) except that we did not use syntactic relations). We removed the features with occurrence frequency (counted in both official training set and official test set) less than three times.

If the estimated number of senses (or sense number) is more than the number of senses in the initial tagged corpus X_L, then the results from order identification based methods will consist of the instances from new classes. When assessing the agreement between these classification results and the ground-truth results on official test set, we will encounter the problem that there is no sense tag for each instance in new classes. Slonim and Tishby (2000) proposed to assign documents in each cluster with the most dominant class label in that cluster,

[12] Here the accuracy is as same as the precision measure in SENSEVAL-3, and the recall of all the methods we evaluated is 100% since we attempted to label all the instances in test data.

and then conducted evaluation on these labeled documents. Here we will follow their method for assigning sense tags to new classes from LPU, clustering based order identification process, and ELP based order identification process. We assigned the instances from new classes with the dominant sense tag in that cluster. The result from LPU always includes only one new class. We assigned the instances from the new class with the dominant sense tag in that cluster. When all instances have their sense tags, we evaluated the results using the accuracy on official test set in SENSEVAL-3.

3.2. Results on sense disambiguation

Table 4 summarizes the accuracy of SVM, LPU, the semi-supervised *k*-means clustering algorithm with correct sense number $|S|$ or estimated sense number $\hat{k}_{X_{\mathrm{L+U}}}$ as input, and the ELP algorithm with correct sense number $|S|$ or estimated sense number $\hat{k}_{X_{\mathrm{L+U}}}$ as input using various incomplete tagged data. The bottom row in Table 4 lists the average accuracy of each algorithm over six experimental settings. Using $|S|$ as input means that we do not perform order identification procedure, while using $\hat{k}_{X_{\mathrm{L+U}}}$ as input is to perform order identification and obtain the classification results on X_{U} at the same time.

Given the correct sense number as input, the average accuracy of the ELP algorithm and the clustering algorithm are 52.5% and 46.1%, respectively. When using the estimated sense number as input, the ELP algorithm and the clustering algorithm achieved 48.9% and 43.8% as the average accuracy. We can see that the ELP based method outperforms the clustering based method in terms of average accuracy under the same experiment setting, e.g. using the correct sense number as input or the estimated sense number as input. Moreover, using the correct sense number as input helps to improve the overall performance of both clustering based method and ELP based method. The two methods, the ELP based method and the clustering based method, outperform the other two algorithms, SVM and LPU.

Comparing the performance of the same system with different tagged datasets (from the first experiment to the third experiment, and from the fourth experiment to the sixth experiment), we can see that in most of cases, the performance of SVM, LPU, the ELP based method, and the clustering based method was improved when using more labeled data. For example, when $S_{\mathrm{subset}} = s_1$ (the most frequent sense is missing), the accuracy of SVM, LPU, the clustering based method with $|S|$ as input, and the ELP based method with $|S|$ as input, the clustering based method with $\hat{k}_{X_{\mathrm{L+U}}}$ as input, and the ELP method with $\hat{k}_{X_{\mathrm{L+U}}}$ as input are 30.6%, 22.3%, 43.9%, 47.8%, 40.0%, and 38.7%. The performance of these methods were improved to 67.0%, 53.4%, 48.7%, 67.2%, 52.4%, and 69.1% when $S_{\mathrm{subset}} = s_3$ (only a rare sense is missing). Furthermore, ELP based method outperforms other methods in terms of accuracy when rare senses (e.g. s_3) are missing in the

Table 4
This table summarizes the accuracy of SVM, LPU, the semi-supervised *k*-means clustering algorithm with correct sense number $|S|$ or estimated sense number $\hat{k}_{X_{\mathrm{L+U}}}$ as input, and the ELP algorithm with correct sense number $|S|$ or estimated sense number $\hat{k}_{X_{\mathrm{L+U}}}$ as input on official test data of English Lexical Sample task in SENSEVAL-3 when given various incomplete tagged datasets

	SVM (%)	LPU (%)	The clustering algorithm with $\|S\|$ as input (%)	The ELP algorithm with $\|S\|$ as input (%)	The clustering algorithm with $\hat{k}_{X_{\mathrm{L+U}}}$ as input (%)	The ELP algorithm with $\hat{k}_{X_{\mathrm{L+U}}}$ as input (%)
$S_{\mathrm{subset}} = \{s_1\}$	30.6	22.3	43.9	47.8	40.0	38.7
$S_{\mathrm{subset}} = \{s_2\}$	59.7	54.6	44.0	62.4	48.5	62.6
$S_{\mathrm{subset}} = \{s_3\}$	67.0	53.4	48.7	67.2	52.4	69.1
			$S_{\mathrm{subset}} = \{s_1, s_2\}$	14.6	13.1	44.4
	35.6	33.0				
40.2						
			$S_{\mathrm{subset}} = \{s_1, s_3\}$	25.7	21.1	48.5
	39.8	31.0				
37.9						
			$S_{\mathrm{subset}} = \{s_2, s_3\}$	56.2	53.1	47.3
	46.6	58.7				
59.4						
Average accuracy	42.3	36.3	46.1	52.5	43.8	48.9

Table 5
These two tables provide the mean and standard deviation of absolute values of the difference between ground-truth results $|S|$ and sense numbers estimated by clustering or ELP based order identification procedure respectively

	The clustering based method	The ELP based method
$S_{\text{subset}} = \{s_1\}$	1.3 ± 1.1	2.2 ± 1.1
$S_{\text{subset}} = \{s_2\}$	2.4 ± 0.9	2.4 ± 0.9
$S_{\text{subset}} = \{s_3\}$	2.6 ± 0.7	2.6 ± 0.7
$S_{\text{subset}} = \{s_1, s_2\}$	1.2 ± 0.6	1.6 ± 0.5
$S_{\text{subset}} = \{s_1, s_3\}$	1.4 ± 0.6	1.8 ± 0.4
$S_{\text{subset}} = \{s_2, s_3\}$	1.8 ± 0.5	1.8 ± 0.5

tagged data. It seems that the ELP based method has the ability to find rare senses with the use of tagged and untagged corpora.

The LPU algorithm can deal with only one-class classification problem. Therefore the labeled data of other classes cannot be used when determining the positive labeled data for current class. The ELP algorithm can use the labeled data of all the known classes to determine the seeds of new classes. It may explain why LPU does not outperform ELP although LPU can correctly estimate the sense number in $X_{\text{L+U}}$ when only one sense is missing in X_{L}.

When only few labeled examples are available, the noise in labeled data makes it difficult to learn confidence score (each entry in $Y_{D_{\text{U}}}$). Therefore using the classification confidence criterion may lead to poor performance of seed selection for new classes if confidence score is not accurate. It may explain why ELP based method does not outperform clustering based method with small labeled data (e.g. $S_{\text{subset}} = \{s_1\}$).

3.3. Results on sense number estimation

Table 5 provides the mean and standard deviation of absolute difference values between ground-truth results $|S|$ and sense numbers estimated by the clustering or ELP based order identification procedures respectively. For example, if the ground truth sense number of word w is k_w, and the estimated value is $\hat{k}_w$, then the absolute value of the difference between these two values is $|k_w - \hat{k}_w|$. Therefore we can have this value for each word. Then we calculated the mean and deviation on this array of absolute values. LPU does not have the order identification capability since it always assumes that there is one new class in unlabeled data, and does not further differentiate the instances from these new classes. Therefore we do not provide the order identification results of LPU.

From the results in Table 5, we can see that the estimated sense numbers are closer to the ground truth results when using less labeled data. Moreover, the clustering based method performs better than the ELP based method in terms of order identification when using small labeled data (e.g. $S_{\text{subset}} = \{s_1\}$). It seems that ELP is not robust to the noise in small labeled data, compared with the semi-supervised k-means clustering algorithm.

4. Related work

The work closest to ours is partially supervised classification or building classifiers using positive and unlabeled examples, which has been studied in machine learning community (Denis et al., 2002; Liu et al., 2003; Manevitz and Yousef, 2001; Yu et al., 2002). They try to learn from positive and unlabeled examples for a given class. After the learning procedure, unlabeled examples will be classified into positive examples and negative examples. These methods always assume the occurrence of negative examples in unlabeled data. Moreover, they cannot group negative examples into meaningful clusters. In contrast, our algorithm can find the occurrences of new senses (represented by negative examples) and further differentiate the new senses by grouping those negative examples into clusters. Semi-supervised clustering (Wagstaff et al., 2001) may be used to perform classification by the use of labeled and unlabeled examples, but it encounters the same problem of partially supervised classification that model order cannot be automatically estimated.

Levine and Domany (2001) and Lange et al. (2002) proposed cluster validation based criteria for cluster number estimation. However, they showed the application of the cluster validation method only for unsupervised learning. Our work can be considered as an extension of their methods in the setting of partially supervised learning by incorporating a partially supervised classification algorithm (the ELP algorithm) in the order identification process.

In natural language processing community, the work that is closely related to ours is word sense discrimination which can induce senses by grouping occurrences of a word into clusters without the use of labeled training data and sense inventories (Fukumoto and Suzuki, 1999; Pedersen and Bruce, 1997; Schütze, 1998).

Schutze's approach firstly selected important contextual words using χ^2 or local frequency criterion. With the χ^2 based criterion, those contextual words whose occurrence depended on whether the ambiguous word occurred were chosen as features. When using local frequency criterion, his algorithm selected top *n* most frequent contextual words as features. Then each context of occurrences of target word was represented by second order co-occurrence based context vector. Singular value decomposition (SVD) was conducted to reduce the dimensionality of context vectors. Then the reduced context vectors were grouped into a pre-defined number of clusters whose centroids corresponded to senses of target word.

Pedersen and Bruce, 1997 described an experimental comparison of three clustering algorithms for word sense discrimination. Their feature sets included morphology of target word, part of speech of contextual words, absence or presence of particular contextual words, and collocation of frequent words. Then occurrences of target word were grouped into a pre-defined number of clusters. Similar with many other algorithms, their algorithm also required the cluster number to be provided.

Fukumoto and Suzuki (1999) presented a term weight learning algorithm for verb sense disambiguation, which can automatically extract nouns co-occurring with verbs and identify the number of senses of an ambiguous verb. The weakness of their method is to assume that nouns co-occurring with verbs are disambiguated in advance and the number of senses of target verb is no less than two.

Another related work is unknown word sense detection by Erk (2006). He addressed the problem of unknown word sense detection as the identification of corpus occurrences that are not covered by a given sense inventory. He modeled this problem as an instance of outlier detection, using a simple nearest neighbor-based approach to measure the resemblance of a new item to a training set. His work has the problem that occurrences with unknown senses cannot be grouped into meaningful clusters even if there are two or more unknown senses in test data.

Word sense discrimination methods use unsupervised methods to solve sense disambiguation problem without the use of labeled training data, while our algorithm can utilize both labeled data and unlabeled data. In comparison with semi-supervised sense disambiguation methods, e.g. bootstrapping (Yarowsky, 1995), our algorithm can try to find missing senses from unlabeled data.

5. Conclusions and future work

In this paper, we present a model order identification based partially supervised classification algorithm, which can classify unlabeled data into positive and negative examples, and further group negative examples into a natural number of clusters. Experimental results on SENSEVAL-3 data for word sense disambiguation task indicate that this classification process with ELP as the base classifier achieves better performance than SVM, LPU, and the classification process with the semi-supervised *k*-means clustering as the base classifier.

In the future, we would like to put our efforts on improvement of the order identification procedure with labeled and unlabeled data. Other possible work includes the investigation of more principled method to estimate the size of seed set in ELP.

References

Brown P., Stephen, D.P., Vincent, D.P., Robert, M., 1991. Word sense disambiguation using statistical methods. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics.

Dagan, I., Itai, A., 1994. Word sense disambiguation using a second language monolingual corpus. Computational Linguistics 20 (4), 563–596.

Denis, F., Gilleron, R., Tommasi, M., 2002. Text classification from positive and unlabeled examples. In: Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems.

Diab, M., Resnik, P., 2002. An unsupervised method for word sense tagging using parallel corpora. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 255–262.

Erk, K., 2006. Unknown word sense detection as outlier detection. In: Proceedings of the Conference of North America Chapter of the Association for Computational Linguistics, NYC, USA.

Fukumoto, F., Suzuki, Y., 1999. Word sense disambiguation in untagged text based on term weight learning. In: Proceedings of the 9th Conference of European Chapter of the Association for Computational Linguistics, pp. 209–216.

Hearst, M., 1991. Noun homograph disambiguation using local context in large text corpora. In: Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora, vol. 24(1), pp. 1–41.

Karov, Y., Edelman, S., 1998. Similarity-based word sense disambiguation. Computational Linguistics 24 (1), 41–59.

Lange, T., Braun, M., Roth, V., Buhmann, J.M., 2002. Stability-based model selection. Advances in Neural Information Processing Systems 15.

Leacock, C., Miller, G.A., Chodorow, M., 1998. Using corpus statistics and WordNet relations for sense identification. Computational Linguistics 24 (1), 147–165.

Lee, Y.K., Ng, H.T., 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In: Proceedings of the 7th Conference on Empirical Methods in Natural Language Processing, pp. 41–48.

Levine, E., Domany, E., 2001. Resampling method for unsupervised estimation of cluster validity. Neural Computation 13, 2573–2593.

Li, H., Li, C., 2004. Word translation disambiguation using bilingual bootstrapping. Computational Linguistics 30 (1), 1–22.

Lin, J., 1991. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory 37 (1), 145–150.

Liu, B., Dai, Y., Li, X., Lee, W.S., & Yu, P., 2003. Building text classifiers using positive and unlabeled examples. In: Proceedings of the Third IEEE International Conference on Data Mining, Melbourne, Florida.

Manevitz, L.M., Yousef, M., 2001. One class SVMs for document classification. Journal of Machine Learning 2, 139–154.

Mihalcea, R., Moldovan, D., 1999. An automatic method for generating sense tagged corpora. In: Proceedings of the 16th National Conference on Artificial Intelligence and Eleventh Conference on Innovative Applications of Artificial Intelligence, Orlando, Florida, USA, pp. 461–466.

Mihalcea, R., 2004. Co-training and self-training for word sense disambiguation. In: Proceedings of the Conference on Natural Language Learning.

Mihalcea, R., Chklovski, T., Kilgariff, A., 2004. The SENSEVAL-3 English Lexical Sample Task. SENSEVAL-2004.

Ng, H.T., Wang, B., Chan, Y.S., 2003. Exploiting parallel texts for word sense disambiguation: an empirical study. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 455–462.

Niu, Z.Y., Ji, D.H., Tan, C.L., 2005. Word sense disambiguation using label propagation based semi-supervised learning. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics.

Park, S.B., Zhang, B.T., Kim, Y.T., 2000. Word sense disambiguation by learning from unlabeled data. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics.

Pedersen, T., 2000. A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation. In: Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics.

Pedersen, T., Bruce, R., 1997. Distinguishing word senses in untagged text. In: Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing, pp. 197–207.

Pham, T.P., Ng, H.T., Lee, W.S., 2005. Word sense disambiguation with semi-supervised learning. In: Proceedings of the 20th National Conference on Artificial Intelligence, Pittsburgh, Pennsylvania, USA, pp. 1093–1098.

Schütze, H., 1998. Automatic word sense discrimination. Computational Linguistics 24 (1), 97–123.

Wagstaff, K., Cardie, C., Rogers, S., Schroedl, S., 2001. Constrained *k*-means clustering with background knowledge. In: Proceedings of the 18th International Conference on Machine Learning, pp. 577–584.

Yarowsky, D., 1995. Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, pp. 189–196.

Yu, H., Han, J., Chang, K.C.-C., 2002. PEBL: positive example based learning for web page classification using SVM. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery in Databases.

Zhu, X., Ghahramani, Z., 2002. Learning from labeled and unlabeled data with label propagation. CMU CALD Tech Report CMU-CALD-02-107.